УДК 004.94:658 doi:10.20998/2413-4295.2018.26.31

КО-КЛАСТЕРИЗАЦИЯ ДАННЫХ МНОГОМЕРНЫХ АТРИБУТОВ КАЧЕСТВА ДЛЯ ОЦЕНКИ ФАКТОРОВ ВЗАИМНОГО ВЛИЯНИЯ

С. В. ШТАНГЕЙ[1], И. В. ТЕРЕЩЕНКО[1*], А. И. ТЕРЕЩЕНКО[2]

[1] кафедра информационно-коммуникационной инженерии, Харьковский Национальный Университет Радиоэлектроники, Харьков, УКРАИНА
[2] кафедра управления информационной и кибернетической безопасностью, Государственный Университет Телекоммуникаций, Киев, УКРАИНА
[*]email: iter@ukr.net

АННОТАЦИЯ *В статье предлагается метод ко-кластеризации стохастических данных многомерных критических параметров процесса (CPPs) с целью оценки влияния обнаруженных факторов на многомерные атрибуты критического качества (CQAs) продукта на стадии первоначального проектирования процесса производства. Метод представляет новый подход к обеспечению качества продукта, который учитывает проблему ко-кластеризации массивов данных CPPs для определения каузальной связи с CQAs. Используется технология неметрического многомерного шкалирования (NMDS) для определения исходных параметров ко-кластеризации.*
Ключевые слова: *качество через дизайн; критические атрибуты качества; ко-кластеризация; многомерный статистический анализ*

CO-CLUSTERING THE DATA OF MULTIVARIATE QUALITY ATTRIBUTES FOR THE ESTIMATION OF MUTUAL INFLUENCE FACTORS

S. SHTANGEY[1], I. TERESHCHENKO[1*], A. TERESHCHENKO[2]

[1] department of info-communication engineering, Kharkiv National University of Radio Electronics, Kharkiv, UKRAINE
[2] department of management of information and cyber security, State University of Telecommunications, Kyiv, UKRAINE

ABSTRACT *Nowadays, competitiveness and efficiency of companies must be continuously improved to face worldwide competitors. Evolutions of computer-intensive technologies development and massive integration of numerical simulations to the design process require new methodologies for numerical design of experiments which use to improve quality of products by taking into account uncertainties in product development. Simultaneous clustering of rows and columns, known as co-clustering, is an important method of two-way analysis of empirical data for practical approaches. The article propose a method of contingency data co-clustering based on multivariate statistical analysis (MSA) for evaluating the influence of critical process parameters (CPPs) factors on the time multivariate critical quality attributes (CQAs) of product. Factorized multivariate CPPs increase the possibility to use methods of multivariate statistical analysis for evaluating the influence of CPPs on the multivariate CQAs. The article solves the objective of product's quality assurance at the stage of the initial manufacture process design in accordance with the process-analytical technology for the design of modern certified manufacturing known as «quality-by-design» (QbD). The method proposed in the article presented a new approach of product's quality assurance which takes into account the block clustering problem on both the individuals and variables parameters for data arrays of computer format. A key feature of the article is the use of nonmetric multidimensional scaling (NMDS) technology to determine the initial parameters of co-clustering. Cluster analysis is an essential tool in different kinds of scientific areas including data mining. Therefore, the subject matter of the article is relevant and given for further development.*
Keywords: *quality-by-design; critical quality attributes; co-clustering; multivariate statistical analysis*

Введение

Одно из важнейших условий развития любого современного производства – обеспечение постоянно высокого сертифицированного качества конечного продукта. Быстро меняющаяся конъюнктура рынка и усиление конкуренции объективно способствуют имплементации высокотехнологичных методов обеспечения качества на этапе разработки продукции [1-3] с дальнейшим менеджментом всего производственного процесса в соответствии с определёнными критическими атрибутами качества CQA (Critical Quality Attributes) [4].

Современная идеология использования высокотехнологичных аналитических/синтетических методов при достижении стандартов качества на ранних этапах проектирования [1] как многоцелевая задача [5] имплементирована в концепции QbD (Quality-by-Design) [4, 6, 7], которая активно внедряется FDA (United States Food Drug Administration) [4] и получает распространение в разных отраслях экономики. В качестве важного инструмента QbD рассматривается аналитическая технология PAT (Process Analytical Technology) [4, 8], которая оперирует информационными технологиями многомерного статистического анализа (MSA, Multivariate Statistical Analysis) для обеспечения

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

качества продукции и поддержания его критических атрибутов в пределах установленных стандартами значений пространства разработки (design space). Методология QbD базис качества продукта закладывает на стадии проектирования и вводит понятие «пространства разработки» (design space) как некую комбинацию одного или нескольких параметров (атрибутов) процесса, влияющих на желаемое свойство (качество) продукта [6, 9]. Такой подход формирует требования к концепции PAT рассматривать контроль качества продукта как научное направление, цель которого снизить риск для пациентов/потребителей путем контроля производства на основе глубокого понимания процесса [4].

С точки зрения PAT, процесс считается хорошо понятым, когда [4]:
– Все критические источники изменчивости идентифицируются и объясняются;
– Изменчивость управляется процессом;
– Атрибуты качества продукта могут быть точно и надежно предсказаны.

В качестве современного инструментария PAT используются методы прикладной хемометрики [10] и компьютерно-интенсивные информационные технологии [11, 12], которые могут быть отнесены к четырем классам [4]:
– Многомерные инструменты для проектирования, сбора и анализа данных;
– Анализаторы процессов;
– Инструменты управления технологическим процессом;
– Непрерывное совершенствование и инструменты управления знаниями.

Многомерные данные многофакторного эксперимента и результаты их анализа позволяют установить допустимые диапазоны изменчивости каждого критического параметра производственного процесса CPPs (Critical Process Parameters) исходя из степени влияния на атрибуты критического качества продукта (CQAs) или ожидаемую спецификацию качества продукта [6].

Эффективным методом дескриптивного и индуктивного статистического оценивания каузальных связей $CPPs \rightarrow CQAs$ является многомерный кластерный анализ [13]. Современные методы кластерного анализа предлагают блочные модели классификации разных видов данных в т. ч. модели исследования смеси распределений многомерных массивов (матриц, таблиц) данных, которые позволяют установить латентные зависимости параметров наблюдений [14-18].

Данные технологии блоковой кластеризация (бикластеризации) или двухмодальной кластеризации представляют методику data mining, которая позволяет провести одновременную кластеризацию строк и столбцов матрицы данных. Алгоритмы бикластеризации позволяют формировать подмножества строк (кластеры), которые проявляют сходные свойства через подмножество столбцов (бикластеры). Если рассматривать строки как данные наблюдений, то объединение столбцов, т. е. параметров процесса в однородные группы раскрывает скрытые связи между ними, которые важны для интерпретации в практических приложениях.

В работах [13, 15-17] представлен большой обзор литературы по данной тематике. Новые публикации совершенствуют теорию определения исходных данных для блочного моделирования [19, 20]. Однако не достаточно освещённой остаётся тематика приложения результатов исследований к задачам обеспечения качества продукции, которое должно содержать интерпретацию влияния бикластеризованных CPPs на многомерный отклик процесса производства в виде CQAs. Важной практической проблемой также остаются рекомендации выбора влияющих на сходимость начальных приближений параметров итерационных двухмодальных алгоритмов [19].

Представленные в статье результаты развивают направление интерпретации и соответствуют базовым принципами PAT исследования изменчивости показателей производственных процессов и обеспечения качества продукта на ранних стадиях проектирования. В качестве инструментов исследований использованы информационные технологии многомерного блочного кластерного анализа.

Цель работы (цель и задачи исследования)

Цель статьи представить метод многомерного статистического кластерного анализа для оценивания характера и особенностей факторного влияния многомерных массивов временных данных критических параметров производственного процесса CPPs (Critical Process Parameters) на критические атрибуты качества продукта (CQAs). CPPs и CQAs объединим общим названием «атрибуты качества».

Объектом исследования является процесс обеспечения качества продукта на ранних этапах проектирования производства.

Предметом исследований являются информационные технологии оценки факторного влияния кластеризованных CPPs на CQAs.

Важно, что эти технологии рассматриваются в рамках многомерного статистического анализа (MSA), на котором в настоящее время основаны базовые алгоритмы сбора, обработки и анализа данных. Методы MSA рекомендованы современными версиями мировых стандартов ISO, GMP как инструменты обеспечения качества и безопасности продукции [21-23]. Результаты MSA характеризуют определённый этап процесса производства в соответствии с процессной технологией структурирования функции качества QFD (Quality Function Deployment) [24]. Применение QFD даёт

хорошо интерпретируемые схемы и матрицы данных на каждом из этапов QFD, к которым применимы информационные технологии объектно-ориентированного MSA.

При использовании MSA ставится цель снижения затрат/потерь при стабилизации качества продукта посредством сбора и последующего статистического анализа данных о параметрах процесса на этапе проектных разработок [25]. При этом решается, как минимум, две важнейшие задачи [6]:

1. Определение критических атрибутов качества CQAs (Critical Quality Attributes) как профильные показатели или характеристики, которые необходимо контролировать (прямо или косвенно) для обеспечения качества продукта.

2. Определение характера влияния критически важных параметров процесса CPPs на критические атрибуты качества продукта CQAs, по которому они подразделяются на три класса: неклассифицированные, критические или некритические.

Зависимость CQAs от CPPs рассматривается как функция качества: $CQAs = F(CPP_i)$. Данная функция структурирована соответственно основным этапам производственного/технологического процесса QFD [24], для которых определены критические атрибуты качества и применимы принципы QbD.

Отметим, что гипотеза независимого влияния каждого из компонент многомерных CPPs на критические атрибуты качества продукта CQAs представляет идеалистический подход и, как правило, не является адекватной. В этой связи значение приобретает классификация атрибутов CPPs как факторов объектного (группового) влияния на атрибуты CQAs с дальнейшим многомерным статистическим анализом этого влияния. Данная проблема исследования и регулирования каузальной связи $CPPs \to CQAs$ является актуальной т.к. сопровождает производства на протяжении всего жизненного цикла продукции.

Изложение основного материала

Современные сложные производственные процессы характеризуются связанными ансамблями данных X и Y, включающих определённое количество многомерных компонент, что определяет особенности проведения статистического анализа.

Критические атрибуты выполнения важнейших (критических) процессов представляют собой результаты мониторинга $CPPs - X$ и $CQAs - Y$. В общем случае случайные значения данных параметров (случайных величин) сформированы в массивы (таблицы) временных данных компьютерного формата от аппаратно-программных средств контроля производства. Примем, что задана матрица случайных данных $X = \{x_{ij}; i \in I, j \in J\}$, где I ($i = 1,...,n$) – набор из n объектов (строки матрицы данных, наблюдения), J ($j = 1,...,d$) – набор из d переменных (столбцы матрицы, атрибуты).

В терминах MSA компоненты $X - x_{ij}$ являются предикторами, экзогенными или объясняющими переменными и рассматриваются как значения многомерной случайной величины X. Переменным X однозначно соответствует многомерный отклик процесса Y. Компоненты $Y - y_{iu}$ ($i = 1,...,n$; $u = 1,...,b$) являются эндогенными, зависимыми переменными. Категория $X \to Y$ описывается функцией качества процесса производства: $CQAs = F(CPP_i)$

В статье представлена последовательность действий проведения блочной кластеризации массива данных X с целью дальнейшей интерпретации влияния структурированных данных на многомерный отклик процесса производства Y.

Цель блочной кластеризации (бикластеризации) – разделить массив значений матрицы на однородные блоки, где каждый из наборов I и J представлены g и m кластерами соответственно.

Основную идею методов кластеризации можно сформулировать как организацию перестановок объектов и переменных для создания структуры соответствия в виде блоков $g \times m$ массива данных $n \times d$ X.

Существуют алгоритмы, которые выполняют блочную кластеризацию таблиц (матриц) различных типов данных: двоичных, случайных, непрерывных или категориальных [26].

Эти алгоритмы состоят в оптимизации по некоторому критерию некоторой целевой функции от параметров модели, вид которой зависит от типа и особенностей данных X.

Статистические свойства массива данных, тенденции к кластеризации предложено оценивать на этапе разведочного анализа данных (exploratory data analysis) с помощью объектно-ориентированного языка R [27]. Так, актуальные пакеты языка R используют латентную блочную модель LBM (latent block model) данных с оценкой параметров блочных (кластерных) распределений по EM-алгоритму (expectation-maximization algorithm) [26]. LBM задаёт плотность распределения на множестве X в виде (1):

$$p(X,\mathrm{Q}) = \sum_{(z,w)\in Z\times W} p(z,\mathrm{Q})p(w,\mathrm{Q})f(X \mid z,w;\mathrm{Q}), \quad (1)$$

Распределение (1) при условии независимости разбиений по строкам и столбцам: $row\mathbf{z} = (z_{ik})_{i,k}$

$column\mathbf{w} = (w_{jl})_{jl}$, $k = 1,\ldots,g$, $l = 1,\ldots,m$,

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

а также блочной специфичностью распределений $f(X \mid z, w; Q)$ может быть записано в виде (2):

$$p(X, Q) = \sum_{(z,w)\in Z\times W} \prod_{ik} \pi_k^{z_{ik}} \prod_{jl} p_l^{w_{jl}} \prod_{ijkl} f(x_{ij}, \alpha_{kl})^{z_{ik} w_{jl}}, \quad (2)$$

где Z и W – множества всех возможных кластеров z из I и w из J, $f(x_{ij}, \alpha_{kl})^{z_{ik} w_{jl}}$ – плотность вероятности блочных распределений x_{ij} с параметрами α_{kl}. Q – вектор параметров $Q = (\pi, p, \alpha)$, где $\pi = (\pi_1, ..., \pi_g)$ и $\rho = (\rho_1, ..., \rho_m)$ вероятности того, что строка и столбец относятся к компонентам k-й строки и l-го столбца соответственно, $\alpha_{kl} \in \alpha$.

Важно, что при блочной кластеризации матрица X размером $n \times d$ трансформируется в матрицу весов α_{kl} x^{zw} размером $q \times m$, что можно трактовать как сжатие данных. Кластеризация по строкам и столбцам даёт взвешенные значения атрибутов (d) и наблюдений (n) в образованных классах, которые задаются матрицами $x^z (g \times d)$ и $x^w (n \times m)$. Процесс блочной кластеризации и сжатие данных иллюстрирует рис.1.

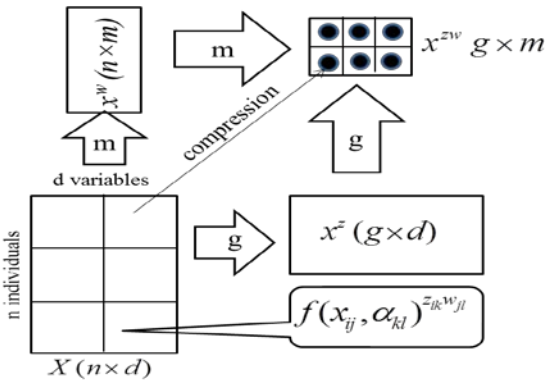


Рис. 1 – Процесс блочной кластеризации и сжатие данных

Для модели распределения (2) полными данными считаются вектор (X, z, w), где ненаблюдаемыми параметрами z и w являются кластеры, раскрывающие скрытые связи данных массива X. Оценки вектора параметров Q находят по максимуму логарифма правдоподобия (3):

$$L(Q; X, z, w) = \sum_{ik} z_{ik} \log \pi_k + \sum_{jl} w_{jl} \log p_l + \sum_{ijkl} z_{ik} w_{jl} \log f(x_{ij}, \alpha_{kl}), \quad (3)$$

Для максимизации (3) используется итерационный неиерархический EM-алгоритм, который предполагает наличие функции плотности вероятности для каждого кластера с соответствующим значениями математического ожидания и дисперсии или/и вектора параметров Q [26].

На *E-шаге (expectation)* вычисляется ожидаемое значение функции правдоподобия, при этом текущее приближение скрытых переменных вектора Q рассматривается как наблюдаемое.

На *M-шаге (maximization)* вычисляется оценка Q максимизирующая правдоподобие (3) и таким образом уточняется значение Q, вычисляемое на E-шаге. Затем это новое значение Q используется для E-шага на следующей итерации – c (4). Алгоритм выполняется до сходимости.

$$Q(Q, Q^c) = \sum_{ik} s_{ik}^c \log \pi_k + \sum_{jl} t_{jl}^c \log p_l + \sum_{ijkl} e_{ikjl}^c \log f(x_{ij}, \alpha_{kl}), \quad (4)$$

где $t_{ik}^c = P(z_{ik} = 1 \mid x, Q^c), r_{jl}^c = P(w_{jl} = 1 \mid x, Q^c),$
$e_{ikjl}^c = P(z_{ik} w_{jl} = 1 \mid x, Q^c).$

Важно отметить, что рассмотренная схема блочной кластеризации не являются жесткой и допускает комбинирование нескольких процедур различных алгоритмов. В частности, для EM-алгоритма решения прикладных задач распространены модели смешанного Гауссова (Gaussian mixture model), Crobin (Crobin model) и другие виды представления данных [15, 17]. Использование той или иной модификации зависит от конкретного приложения поставленной задачи и адекватности описания эмпирических данных.

Предложена следующая схема бикластеризации массива данных предикторов X с целью дальнейшей оценки влияния на массив многомерных откликов процесса Y:

1. Проведение разведочного анализа данных X с целью определения статистических свойств массива данных и тенденций к кластеризации. На этом шаге также оценивается число возможных кластеров по переменной числа наблюдений n (кластеризация строк).

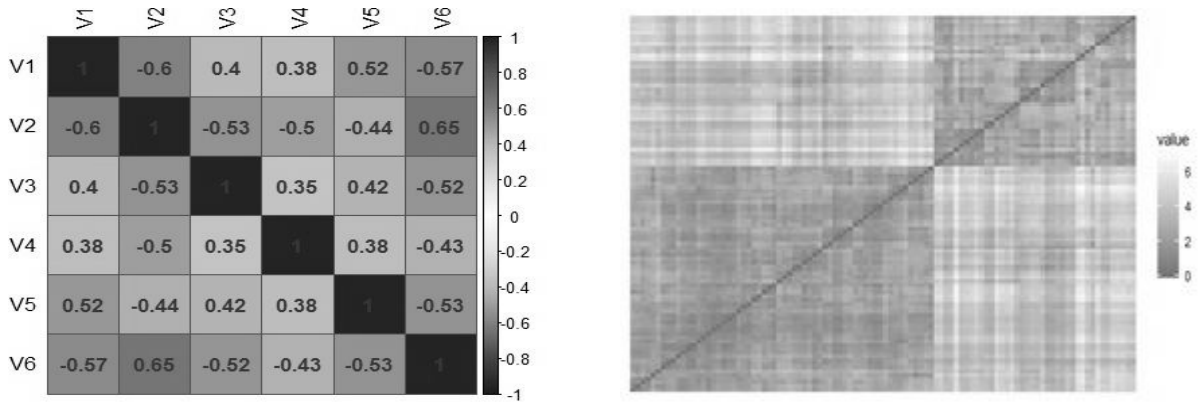
2. Проведение канонического анализа зависимости изменчивости значений матрицы многомерных откликов Y от влияния предикторов X методом неметрического многомерного шкалирования (NMDS, nonmetric multidimensional scaling). На этом шаге число возможных кластеров по переменной предикторов d (кластеризация столбцов) предложено оценивать по степени корреляции этих компонент с осями ординации NMDS.

3. Проведение бикластеризации массива данных предикторов X с оценочными значениями

числа возможных кластеров по переменным наблюдений и предикторов. Анализ и интерпретация полученных результатов.

Исследования выполнены с использованием программных пакетов языка R.

Разведочный анализ свойств предикторов производился по корреляционной матрице и визуализации тенденции образования кластеров VAT (Visual Assessment of cluster Tendency), как показано на рис. 2a,b .

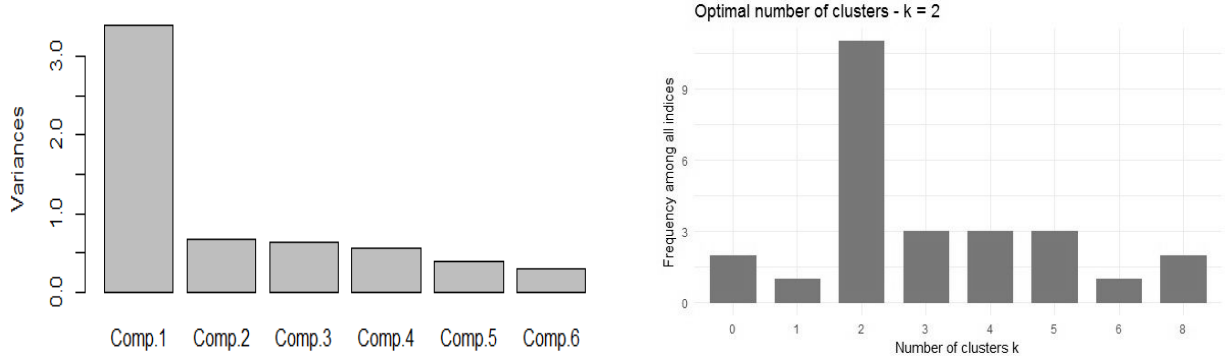


a *b*

Рис. 2 – Корреляционная матрица компонент предикторов X – a и визуализация тенденций образования кластеров – b

Потенциальное объединение наблюдений в группы представлено темными квадратами вдоль главной диагонали «VAT-диаграммы» рис. 2b.

Оценка веса компонент *X* (variable importance) в общем разбросе данных с помощью метода главных компонент PCA (principle components analysis) и результат объединения предикторов *X* в кластеры, показаны на рис. 3a,b.



a *b*

Рис. 3 – Результат оценки веса компонент предикторов X – a. Результат определения оптимального числа кластеров по переменной числа наблюдений n – b

Поиск оптимальной схемы объединения предикторов *X* в кластеры производился перебором различных комбинаций числа групп, метрик дистанций и методов кластеризации с помощью 30 индексов качества. Оптимальное число кластеров по параметру *n* равно двум, как показано на рис. 3b.

Показатели корреляций (рис.2a) и отсутствие мультиколлинеарности между предикторами, определённые с помощью функций findCorrelation и findLinearCombos языка R, свидетельствуют о превалирующем вкладе случайной составляющей в разброс значений массива *X*.

Влияние изменчивости компонент *X* на компоненты многомерных откликов *Y* исследовалось методом неметрического многомерного шкалирования (NMDS) этих массивов данных. Проекция значений массива Y на оси ординации NMDS1-NMDS2 с наложением компонент предикторов (ординационный триплот) показаны на рис. 4. Метод NMDS характеризуется наиболее адекватными оценками для больших матриц данных с значительным вкладом случайных факторов в рассеяние значений, что обосновывает его применение в рассматриваемом случае.

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

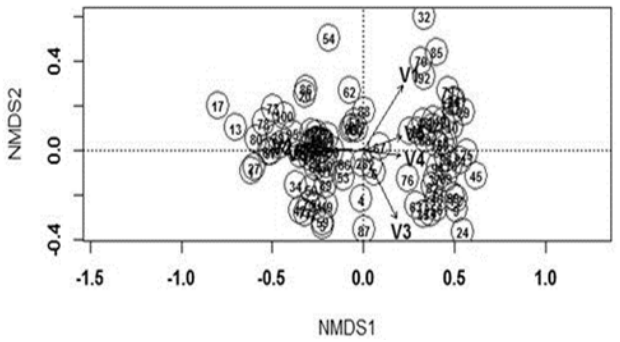


Рис. 4 – Результаты ординации значений откликов Y на оси NMDS1-NMDS2 и связь с компонентами предикторов X

Составляющие предикторов X (V_i) на рис. 4 наложены на плоскость неметрического шкалирования многомерных откликов Y в зависимости от рассчитанных коэффициентов корреляции с осями ординации NMDS1-NMDS2. Из рис. 4 видно, что модель NMDS визуализирует представление о влиянии переменных предикторов X (V_i) на отклики Y. Численной характеристикой являются коэффициенты корреляции каждой из компонент предикторов с осями ординации NMDS1-NMDS2 и оценка статистической значимости этих коэффициентов. Важно отметить, что классы Y хорошо различимы на фоне осей составляющих X, компактное группирование которых на плоскости NMDS1-NMDS2 может говорить о наличии скрытой связи при влиянии на отклики Y и возможности образования кластеров. Технология NMDS даёт возможность исследовать различия не только двух и более матриц данных, но и параметров одной матрицы параметров. Поэтому предложено использовать метод NMDS для определения меры сходства между компонентами массива предикторов X и определения числа кластеров по атрибутивным переменным. Предположение о возможности образования кластеров по переменной предикторов d подтверждается построением ординационного триплота по значениям матрицы предикторов X, рис.5.

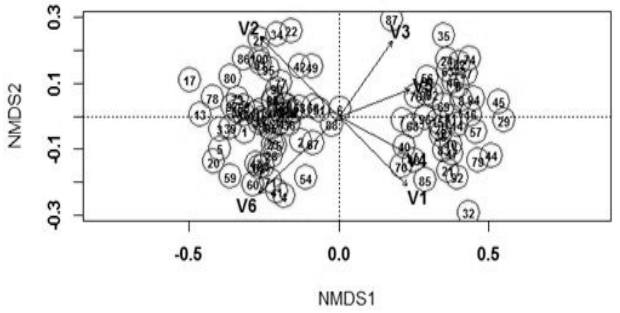
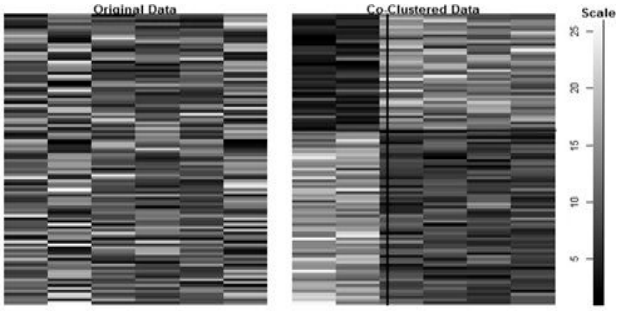


Рис. 5 – Ординационный триплот значений предикторов X на оси NMDS1-NMDS2

Бикластеризация массива X проводилась с использованием программного пакета «blockcluster» языка R для случайных значений массива X распределённых по закону Пуассона [26].

Рисунок 6 визуализирует итог применения инструментов пакета «blockcluster» и образование блоков кластеров по переменной наблюдений n и атрибутов d .



a *b*

Рис. 6 – Результат объединения массива предикторов X – a в блочные кластеры – b

Результаты бикластеризации дополнены блочными и смешанными гистограммами распределений данных массива предикторов X. Программный пакет предоставляет возможность проводить блочную кластеризацию массивов данных как выборок из распределения соответствующего смешанной модели Гаусса.

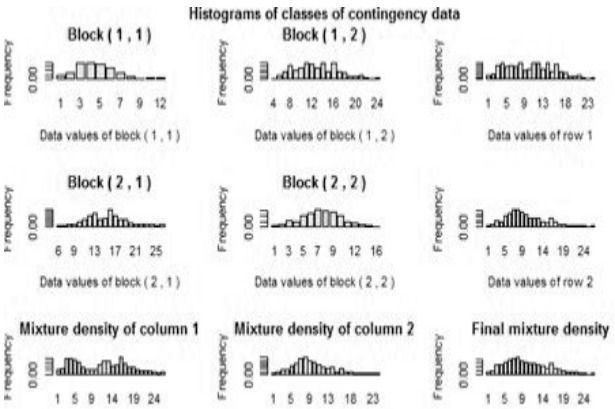


Рис. 7 – Блочные и смешанные гистограммы распределений данных массива предикторов X

Приведенные на рис.6 и рис.7 результаты блочной кластеризации данных массива предикторов X обосновано согласуются с результатами NMDS т.е. ординационный триплот NMDS также как и бикластерная диаграмма дают адекватное представление о классовой структуре исследуемых данных. Оценки NMDS могут быть вполне дополнены последующим анализом блочного деления с помощью бикластерных методов.

Обсуждение результатов

В статье представлены результаты исследования на этапе проектирования процесса производства, в ходе которого проводился многомерный статистический анализ причинно-следственных связей между изменчивостью компонент отклика процесса и действием определенных факторов, отождествляемых с технологическими параметрами процесса.

Факторы получены как результат блочного группирования наблюдений массива объясняющих переменных в кластеры для разных метрик оценивания разброса значений в многомерном пространстве. Кластеры предложено трактовать как факторы влияния критических параметров процесса производства на критические атрибуты качества продукта.

Исследование содержит элементы разведочного, описательного и аналитического статистического анализа.

Разведочная характеристика данных использует показатели корреляционных свойств и тенденций образования кластеров исследуемого массива данных (рис. 2a,b). Кластеры детализированы и выделены темными квадратами вдоль главной диагонали «VAT-диаграммы» рис. 2b.

Значения относительного вклада каждой компоненты предикторов (Cmp.m) в общий разброс данных были получены с помощью метода PCA и сведены в табл. 1.

Таблица 1 – Вклад компонент предикторов в общий разброс данных массива X

Cmp Prmtrs	Cmp 1	Cmp 2	Cmp 3	Cmp 4	Cmp 5	Cmp 6
Stndrd devtn	1.84	0.81	0.79	0.74	0.62	0.55
Prprtn of Vrnce	0.57	0.11	0.10	0.09	0.06	0.05
Cmltv Propor -tion	0.57	0.68	0.78	0.88	0.94	1.00

Отметим, что все компоненты многомерного массива X за исключением Cmp.1 (57%) характеризуются практически равным вкладом (proportions of variance близко к 10%) в разброс значений. Доступная в этом тесте тенденция кластеризации характеризуется статистикой Хопкинса (The Hopkins statistic, 0.334) показывает умеренную склонность к образованию кластеров при весомом влиянии случайного разброса данных.

Поскольку предлагаемые методы бикластеризации используют итерационные процедуры критичные к выбору начального приближения числа разбиений массивов данных по параметрам наблюдений и атрибутов [19, 26], предложено оценивание возможного числа блоков проводить по результатам разведочного анализа данных и применения метода NMDS. Далее показано, что метрика неметрического многомерного шкалирования даёт адекватные оценки группирования для матриц данных с существенным вкладом стохастичности в разброс значений.

Так, ожидаемое число разбиений по переменной наблюдений можно получить посредством сравнительных процедур перестановочных тестов для различных сочетаний числа групп и метрик дистанций (рис.3b).

Отметим, что ординация составляющих X на плоскости компонент NMDS потенциально оценивает количество блоков по переменной атрибутов (рис.4). При этом важно подчеркнуть, что классификация атрибутов проводится именно по ординационному триплоту категории $X \to Y$ характеризующему каузальную связь массива атрибутов X с многомерными откликами Y.

В табл. 2 приведены коэффициенты корреляции компонент предикторов V_i с осями ординации NMDS1-NMDS2 и статистическая значимость (p) этих коэффициентов.

Имеет место высокая корреляция составляющих предикторов с осями ординации NMDS и существенная статистическая значимость коэффициентов корреляции (p=0.01).

Таблица 2 – Статистическая значимость (p) коэффициентов корреляции компонент X (V_i) с осями ординации NMDS

NMDS V	NMDS1	NMDS2	r2	Pr(>r)
V1	0.59247	0.80559	0.8215	0.001
V2	-0.99910	0.04248	0.8738	0.001
V3	0.50809	-0.86131	0.7751	0.001
V4	0.99283	-0.11953	0.2704	0.001
V5	0.95839	0.28547	0.2925	0.001
V6	-0.99975	-0.02231	0.5143	0.001

В статье акцентировано внимание, что эти закономерности ординационного триплота NMDS далее позволяют определить число кластеров по переменной атрибутов массива X и связать концентрацию разброса значений с составом компонент атрибутов (рис. 5).

Количественные показатели дескриптивных статистических исследований (кластеры $g \times m$) позволяют провести блочную кластеризацию массива X (рис.6). С помощью программного пакета «blockcluster» языка R оригинальные данные матрицы X распределённые по закону Пуассона группируются в блочную таблицу 2×2.

Вид блочных гистограмм распределений значений предикторов на рис. 7 в основном

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

унимодален с наличием достаточно больших частот некоторых величин, что как уже отмечалось объясняется влиянием случайных факторов. Значения строковых и столбцевых блочных вероятностей приведены в табл. 3.

Таблица 3 – Значения блочных вероятностей компонент строк и столбцов

Row proportions	0.400	0.599
Column proportions	0.666	0.333

Этот фрейм данных раскрывает скрытую связь компонент предикторов как критических параметров процесса при их влиянии на многомерный отклик процесса в виде критических атрибутов качества продукта. Отметим, что предложенная логика статистического анализа в практическом приложении позволяет конкретизировать технологические параметры вариабельности процесса производства на стадии проектирования и снизить возможные потери от снижения качества продукции.

Выводы

Актуальным направлением анализа каузальных связей массивов критических атрибутов качества процесса производства является совместное или комбинированное использование различных статистических методов. Этому во многом способствует объектно-ориентированный подход к исследованиям, реализованный в языке программирования R.

Блочная кластеризация массивов эмпирических данных критических атрибутов качества процесса производства при многомерных статистических исследованиях каузальной связи позволяет определить практически значимые параметры и их латентные зависимости.

При отсутствии аналитической или физической зависимости между составляющими предикторов, образование классов по атрибутивному измерению объясняет изменение откликов как следствие влияния косвенных (не явных) факторов, связанных с технологией процесса.

Важным результатом является заключение о возможности использования метода NMDS для определения тенденции кластеризации и числа возможных кластеров по атрибутивным параметрам исследуемого массива. Оценки бикластеризации релевантны результатам оценок методом NMDS в отношении состава и характера компонентного влияния предикторов на многомерный отклик процесса.

Перспектива дальнейших исследований состоит в синтезе новых или имплементации известных методов MSA для анализа данных технологических процессов на ранних стадиях проектирования с целью достижения высокого качества продукции.

Список литературы

1. **Steimera, C**. Model-based design process for the early phases of manufacturing system planning using SysML / **C. Steimera, J. Fischerb, J. C. Auricha** // *27th CIRP Design 2017. Procedia CIRP*. – 2017. – P. 163-168. – doi: 10.1016/j.procir.2017.01.036.
2. **Blondet, G**. Simulation Data Management for Adaptive Design Of Experiment. A literature review / **G. Blondet, N. Boudaoud, J. Duigou** // *QUALITA' 2015, Nancy, France*. – 2015. [Electronic resource]. URL: https://hal.archives-ouvertes.fr/hal-01149776 (2018.06.19).
3. **Zwier, Marijn P**. Physics in Design: Real-time numerical simulation integrated into the CAD environment / **Marijn P. Zwier, Wessel W. Wits** // *Procedia CIRP 60*. – 2017. – P. 98-103. – doi: 10.1016/j.procir.2017.01.054.
4. **Zhang, L**. Application of quality by design in the current drug development / **L. Zhang, S. Mao** // Shenyang Pharmaceutical University, Wenhua Road, Shenyang, China: *Asian journal of pharmaceutical sciences*. – 2017. – №.103. – P. 1-8. – doi: 10.1016/j.ajps.2016.07.006.
5. **Mohamed, I**. Progressive Modeling: The Process, the Principles, and the Applications / **I. Mohamed** // *Procedia Computer Science*. – 2013. – V. 16. – P. 39-48. – doi: 10.1016/j.procs.2013.01.005.
6. **Lionberger, R. A**. Quality by Design: Concepts for ANDAs / **R. A. Lionberger, S. L. Lee, L. M. Lee, et al.** // *The AAPS Journal*. – 2008. – 10(2). – P. 268-276. – doi: 10.1208/s12248-008-9026-7.
7. **Pramod, K**. Pharmaceutical product development: A quality by design approach / **K. Pramod, M. A. Tahir, N. A. Charoo, et al.** // *International Journal of Pharmaceutical Investigation*. – Jul-Sep 2016. – 6(3). – P. 129-138. – doi: 10.4103/2230-973X.187350.
8. Guidance for Industry PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. [Electronic resource]. URL: http://www.fda.gov/cder/OPS/PAT.htm (2018.06.19).
9. **Cohen, L**. Research Methods in Education. Sixth edition / **L. Cohen, L. Manion, K. Morrison.** Taylor & Francis e-Library, 2007. – 638 p.
10. **Brereton, R. G**. Applied chemometrics for scientists / **R. G. Brereton.** University of Bristol, UK, John Wiley & Sons Ltd, 2007. – 379 p.
11. **Skold, M**. Computer Intensive Statistical Methods / **M. Skold** // *Mathematical Statistics Centre for Mathematical Sciences Lund University*. – 2005, 2nd printing. – 2006. – 133 p. [Electronic resource]. URL: http://www.maths.lth.se/ (2018.06.19).
12. **Ripley, B. D**. Computer-Intensive Statistics / **B. D. Ripley** // APTS 2011–12 lecture material – 2008. – 70 p. [Electronic resource]. URL: http://www.stats.ox.ac.uk/~ripley/APTS2012/APTS-CIS-lects.pdf (2018.06.19).
13. **Berkhin, P**. A Survey of Clustering Data Mining Techniques / **P. Berkhin** // *ResearchGate.* – 2002. – 59 p. – doi: 10.1007/3-540-28349-8_2.
14. **Govaert, G**. Latent Block Model for Contingency Table / **G. Govaert, M. Nadif** // *Communications in Statistics – Theory and Methods*. – 2010 – 39, 3. – P. 416-425. – doi: 10.1080/03610920903140197.
15. **Charrad, M**. Simultaneous Clustering: A Survey / **M. Charrad, M. Ben Ahmed** // *PReMI'11 Proceedings of*

the 4th international conference on Pattern recognition and machine intelligence. – Springer-Verlag Berlin, 2011. – P. 370-375.

16. **Keribin, C**. Estimation and selection for the latent block model on categorical data / [**C. Keribin, V. Brault, G. Celeux,·et al.**] // *Statistics and Computing*. – November 2015. – V. 25, 6. – P. 1201-1216. – doi: 10.1007/s11222-014-9472-2.
17. **Brault, V.** Co-clustering through Latent Bloc Model: a Review / **V. Brault, M. Mariadassou** // *Journal de la Société Française de Statistique*. – 2015. – 156 (3). – P. 120-139. – doi: 10.1111/evo.12770.
18. **Brault, V**. Fast and Consistent Algorithm for the Latent Block Model / **V. Brault, A. Channarond** // *Electronic Journal of Statistics. arXiv:1610.09005v1 [math.ST]*. – 2016. – 22 p.
19. **Schepers, J.** Maximal Interaction Two-Mode Clustering / **J. Schepers** // *Journal of Classification 34*. – 2017. – P. 49-75. – doi.: 10.1007/s00357-017-9226-x.
20. **Celeux, G.** Variable selection in model-based clustering and discriminant analysis with a regularization approach / **G. Celeux, C. Maugis-Rabusseau, M. Sedki** // *Advances in Data Analysis and Classification*. – Springer Verlag, 2018. – 21 p.
21. ISO 9000 Introduction and Support Package: Guidance on the Concept and Use of the Process Approach for management systems. [Electronic resource]. URL: https://www.iso.org/iso/04_concept_and_use_of_the_process_approach_for_management_systems.pdf (2018.06.19).
22. ISO 22000: 2005 Food safety management systems. [Electronic resource]. URL: https://www.iso.org/standard/35466.html (2018.06.19).
23. Food and Drug Administration. Final Report on Pharmaceutical cGMPs for the 21st Century – A Risk Based Approach. [Electronic resource]. URL: http://www.fda.gov/cder/gmp/gmp2004/GMP_finalreport2004.htm (2018.06.19).
24. **Jaiswal, E. S**. A Case Study on Quality Function Deployment (QFD) / **E. S. Jaiswal** // *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)*. – 2012. – V. 3, 6. – P. 27-35. – doi: 10.9790/1684-0362735.
25. **Araujo, F**. Variable selection methods in multivariate statistical process control: A systematic literature review / **F. Araujo, P. Pimentel, F. S. Fogliatto** // Department of Industrial Engineering, Federal University of Rio Grande do Sul, 90035-190 Porto Alegre, RS, Brazil, *Computers & Industrial Engineering*. – 2018. – V. 115. – P. 603-619. – doi: 10.1016/j.cie.2017.12.006.
26. **Bhatia, P**. Blockcluster: An R Package for Model Based Co-Clustering / **P. Bhatia, S. Iovleff, G. Govaert** // *Journal of Statistical Software*. – 2014. – V. VV, II. – 23 p.
27. **Bhatia, P. S**. A tutorial for blockcluster R package Version 4 / **P. S. Bhatia, S. Iovleff** // *CRAN*, 2018. [Electronic resource]. URL: https://cran.r-project.org/web/packages/blockcluster/vignettes/blockcluster_tutorial.pdf (2018.06.19).

Bibliography (transliterated)

1. **Steimera, C., Fischerb, J., Auricha, J. C.** Model-based design process for the early phases of manufacturing system planning using SysML. *27th CIRP Design 2017. Procedia CIRP*, 2017, 163-168, doi: 10.1016/j.procir.2017.01.036.
2. **Blondet, G., Boudaoud, N., Duigou, J.** Simulation Data Management for Adaptive Design Of Experiment. A literature review, by Blondet, Gaëtan, et al. *QUALITA' 2015*, Nancy, France, 2015. Available at: https://hal.archives-ouvertes.fr/hal-01149776 (19 June 2018).
3. **Zwier, M. P., Wits, W, W.** Physics in Design: Real-time numerical simulation integrated into the CAD environment. *Procedia CIRP 60*, 2017, 98-103, doi: 10.1016/j.procir.2017.01.054.
4. **Zhang, L., Mao, S.** Application of quality by design in the current drug development. *Shenyang Pharmaceutical University, Asian journal of pharmaceutical sciences*, 103, 2017, 1-8, doi: 10.1016/j.ajps.2016.07.006.
5. **Mohamed, I.** Progressive Modeling: The Process, the Principles, and the Applications. *Procedia Computer Science*, 16, 2013, 39-48, doi: 10.1016/j.procs.2013.01.005.
6. **Lionberger, R. A., Lee, S. L., Lee, L. M., et al.** Quality by Design: Concepts for ANDAs. *The AAPS Journal*, 10(2), 2008, 268-276, doi: 10.1208/s12248-008-9026-7.
7. **Pramod, K., Tahir, M. A., Charoo, N. A., et al.** Pharmaceutical product development: A quality by design approach. *International Journal of Pharmaceutical Investigation*, 6(3), 2016, 129-138, doi: 10.4103/2230-973X.187350.
8. Guidance for Industry PAT – *A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.* Available at: http://www.fda.gov/cder/OPS/PAT.htm (19 June 2018).
9. **Cohen, L., Manion, L., Morrison, K.** Research Methods in Education. Sixth edition. *Taylor & Francis e-Library*, 2007, 638.
10. **Brereton, Richard G**. Applied chemometrics for scientists. University of Bristol, UK, *John Wiley & Sons Ltd*, 2007, 379.
11. **Skold, M.** *Computer Intensive Statistical Methods*. Mathematical Statistics Centre for Mathematical Sciences Lund University, 2005, 133. Available at: http://www.maths.lth.se/ (2018.06.19).
12. **Ripley, B. D**. Computer-Intensive Statistics. *APTS 2011–12 lecture material*, 2008, Available at: http://www.stats.ox.ac.uk/~ripley/APTS2012/APTS-CIS-lects.pdf (19 June 2018).
13. **Berkhin, P.** A Survey of Clustering Data Mining Techniques. *ResearchGate*, 2002, 59, doi: 10.1007/3-540-28349-8_2.
14. **Govaert, G., Nadif, M.** Latent Block Model for Contingency Table. *Communications in Statistics – Theory and Methods*, 2010, 39, 3, 416-425, doi: 10.1080/03610920903140197.
15. **Charrad, M., Ben Ahmed, M.** Simultaneous Clustering: A Survey. *PReMI'11 Proceedings of the 4th international conference on Pattern recognition and machine intelligence.* Springer-Verlag Berlin, 2011, 370-375.
16. **Keribin, C., Brault, V.** Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 2015, **25**, 6, 1201-1216, doi: 10.1007/s11222-014-9472-2.
17. **Brault, V., Mariadassou, M.** Co-clustering through Latent Bloc Model: a Review. *Journal de la Société Française de Statistique*, 2015, 156 (3), 120-139, doi: 10.1111/evo.12770.
18. **Brault, V., Channarond, A.** Fast and Consistent Algorithm for the Latent Block Model. *Electronic Journal of Statistics. arXiv:1610.09005v1 [math.ST]*, 2016, 22.
19. **Schepers, J.** Maximal Interaction Two-Mode Clustering. *Journal of Classification*, 2017, **34**, 49-75, doi: 10.1007/s00357-017-9226-x.
20. **Celeux, G., Maugis-Rabusseau, C., Sedki, M.** Variable selection in model-based clustering and discriminant

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

analysis with a regularization approach. *Advances in Data Analysis and Classification. Springer Verlag*, 2018, 21.
21. ISO 9000 Introduction and Support Package: *Guidance on the Concept and Use of the Process Approach for management systems*. Available at: https://www.iso.org/iso/04_concept_and_use_of_the_process_approach_for_management_systems.pdf (19 June 2018).
22. ISO 22000: 2005 *Food safety management systems*. Available at: https://www.iso.org/standard/35466.html (19 June 2018).
23. Food and Drug Administration.*Final Report on Pharmaceutical cGMPs for the 21st Century – A Risk Based Approach*. Available at: http://www.fda.gov/cder/gmp/gmp2004/GMP_finalreport2004.htm (19 June 2018).
24. **Jaiswal, E. S.** A case study on quality function deployment (QFD). *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)*, 2012, 3, 27-35, doi: 10.9790/1684-0362735.
25. **Peres, F., Pimentel, A., Fogliatto, F. S.** Variable selection methods in multivariate statistical process control: A systematic literature review. Department of Industrial Engineering, Federal University of Rio Grande do Sul, 90035-190 Porto Alegre, RS, Brazil, *Computers & Industrial Engineering*, 2018, **115**, 603-619, doi: 10.1016/j.cie.2017.12.006.
26. **Bhatia, P., Iovleff, S., Govaert, G.** *Blockcluster: An R Package for Model Based Co-Clustering*. Journal of Statistical Software, 2014, **VV**, II, 23.
27. **Bhatia, P., Iovleff, S.** A tutorial for blockcluster R package Version 4. *CRAN*, 2018, 1-20. Available at: https://cran.r-project.org/web/packages/blockcluster/vignettes/blockcluster_tutorial.pdf (19 June 2018).

Сведения об авторах (About authors)

Штангей Светлана Викторовна – кандидат технических наук, доцент, Харьковский Национальный Университет Радиоэлектроники, доцент кафедры информационно-коммуникационной инженерии; г. Харьков, Украина; e-mail: qwertysv1@ukr.net.

Svitlana Shtangey – Candidate of Technical Sciences (Ph. D.), Docent, Kharkiv National University of Radio Electronics, Associate Professor, Department of info-communication engineering; Kharkiv, Ukraine; e-mail: qwertysv1@ukr.net.

Терещенко Игорь Владимирович – кандидат технических наук, доцент, Харьковский Национальный Университет Радиоэлектроники, доцент кафедры информационно-коммуникационной инженерии; г. Харьков, Украина; e-mail: iter@ukr.net.

Igor Tereshchenko – Candidate of Technical Sciences (Ph. D.), Docent, Kharkiv National University of Radio Electronics, Associate Professor, Department of info-communication engineering; Kharkiv, Ukraine; e-mail: iter@ukr.net.

Терещенко Антон Игоревич – Государственный Университет Телекоммуникаций, аспирант, кафедра управления информационной и кибернетической безопасностью; Киев, Украина; e-mail: vti62@ukr.net.

Anton Tereshchenko – State University of Telecommunications, post-graduate student, department of management of information and cyber security; Kyiv, Ukraine; e-mail: vti62@ukr.net.

Пожалуйста, ссылайтесь на эту статью следующим образом:

Штангей, С. В. Ко-кластеризация данных многомерных атрибутов качества для оценки факторов взаимного влияния / **С. В. Штангей, И. В. Терещенко, А. И. Терещенко** // *Вестник НТУ «ХПИ», Серия: Новые решения в современных технологиях*. – Харьков: НТУ «ХПИ». – 2018. – № 26 (1302). – Т. 2 . – С. 45-54. – doi:10.20998/2413-4295.2018.26.31.

Please cite this article as:

Shtangey, S., Tereshchenko, I., Tereshchenko, A. Co-clustering the data of multivariate quality attributes for the estimation of mutual influence factors. *Bulletin of NTU "KhPI". Series: New solutions in modern technologies*. – Kharkiv: NTU "KhPI", 2018, **26** (1302), 2, 45-54, doi:10.20998/2413-4295.2018.26.31.

Будь ласка, посилайтесь на цю статтю наступним чином:

Штангей, С. В. Ко-кластеризація даних багатовимірних атрибутів якості для оцінки факторів взаємного впливу / **С. В. Штангей, І. В. Терещенко, А. І. Терещенко** // *Вісник НТУ «ХПІ», Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». – 2018. – № 26 (1302). – Т. 2. – С. 45-54. – doi:10.20998/2413-4295.2018.26.31.

АНОТАЦІЯ *У статті пропонується метод ко-кластеризації стохастичних даних багатовимірних критичних параметрів процесу (CPPs) з метою оцінки впливу виявлених факторів на багатовимірні атрибути критичної якості (CQAs) продукту на стадії початкового проектування процесу виробництва. Метод представляє новий підхід до забезпечення якості продукту, який враховує проблему ко-кластеризації масивів даних CPPs для визначення каузального зв'язку з CQAs. Використовується технологія неметричного багатовимірного шкалювання (NMDS) для визначення вихідних параметрів ко-кластеризації.*

Ключові слова: *якість через дизайн; критичні атрибути якості; ко-кластеризація; багатовимірний статистичний аналіз*

Поступила (received) 16.06.2018